DOCUMENT RESUME

ED 191 874 TM 800 494

AUTHOR Christopherson, Steven L.
TITLE A Numerical Scoring Method for Summaries in Prose Learning Research.
PUB DATE 79
NOTE 8p.: Paper presented at the Annual Meeting of the Eastern Educational Research Association (Kiawah Island, SC, February 20-24, 1979).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Essay Tests: Higher Education: Prose: *Scoring Formulas: Test Validity
IDENTIFIERS Efficiency Ratio: Power Ratio

ABSTRACT

A procedure for creating a numerical index which can be used to score and rank summaries of prose passages is described. The quality of a summary is assumed to be a function of two variables: the number of important ideas mentioned, and the number of words used. Scoring involves multiplying two ratios: a power ratio and an efficiency ratio. The product is consistent with the assumption that better summaries are those which include more of the important ideas while using fewer words. Reliability was checked by rescoring all summaries after six months: the reliability of the index across that period was .86. Criterion-related validity was estimated by correlating the scores with subjective judgments made by the six undergraduate judges as to which summaries were better. Construct validity was checked by an experimental procedure which was expected to affect the quality of summaries. The results are promising, and the scoring procedure appears to have good reliability and validity: its value, however, needs to be tested by others. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED191874

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

A Numerical Scoring Method for Summaries in Prose Learning Research

Steven L. Christopherson
Division of Education
The University of Texas at San Antonio
San Antonio, Texas 78285

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
S. Christopherson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 800494

Paper presented at the Eastern Educational Research Association, 1979.

ABSTRACT

Teachers and researchers have long recognized that summaries are attractive measures of comprehension. Unfortunately, the usual mehtods of scoring and ranking summaries suffer from considerable subjectivity. This paper describes a numerical index which reflects the efficiency of a summary in representing the most important information. The reliability of the index across a six month period was .86. There was significant agreement, $p<.003$, between the index and the mean ratings of judges. Additional evidence of validity is presented.

APR 4 1980

Summaries are attractive measures of comprehension because there is no other way to evoke an articulated view of what a reader believes the message in a communication to have been. There is also no confounding due to recognition or to cues as occurs with multiple-choice or other tests of comprehension. The use of summaries has had its problems, however, because of difficulties involved in rating and ranking the reader-produced protocols. In this paper I describe a procedure for creating a numerical index which can be used to score and rank summaries according to their quality, and then I report on evaluations of the reliability and validity of the index.

PROCEDURE

I have assumed that the quality of a summary is a function of at least two variables: one is the number of important ideas mentioned and the other is the number of words used. The scoring procedure involves multiplying two ratios, one which I call a power ratio and the other an efficiency ratio. The power ratio is the number of important ideas present in the summary divided by the number of important ideas present in the original passage. The efficiency ratio is the number of important ideas present in the summary divided by the number of words in the summary. The product is consistent with the assumption that better summaries are those which include more of the important ideas while using fewer words. For example, a summary which mentions nine out of twelve important ideas and which uses a total of forty-five words would have a score of 9/12 x 9/45 = .15; and a summary with the same number of words, but which mentions eleven out of twelve important ideas, would have a score of 11/12 x 11/45 = .22. Higher scores indicate better summaries.

To score the summaries of a passage, the original passage is divided into idea units (clauses and significant phrases) and some of the idea units are identified as being important. Then the number of important ideas and the total number of words in each summary are counted. Finally, the product of the power ratio and the efficiency ratio, described above, is calculated.

Although all of the steps in the procedure can be performed by the same person, independent judgments were included for an evaluation of the reliability and validity of the scoring procedure. The author and a graduate student independently divided passages of about eighty words each into idea units and then discussed and agreed upon a final division. Six undergraduates, randomly selected from a large pool of subjects, independently judged which of the original idea units they believed to be most important, that is, which should be included in a summary. The judges were permitted to consult the original passages. When four or more judges agreed, an idea unit was labeled "important" for the purpose of scoring summaries. Sixty undergraduates had read and summarized one of four randomly distributed passages. The author counted the number of important ideas and the number of words in each summary. A score for each summary was then calculated by multiplying the power ratio and efficiency ratio for the summary.

Reliability was checked by rescoring all summaries after six months. Reliability was also checked by correlating the original scores with more stringent scores. The more stringent scores were created by requiring that at least five out of the six judges agree that an idea unit was important.

Criterion-related validity was estimated by correlating the scores with subjective judgments made by the six undergraduate judges as to which summaries were better. The summaries were all typed on cards and paired according to passage and summary length, with the maximum difference in length for each pair being five words. The paired summaries were then given to the judges, who were instructed that a better summary is one which more fully presents the essential information from the original passage. When four or more judges agreed, a summary was labeled "better" for this test of the validity of the numerical scores.

Construct validity was checked by an experimental procedure which was expected to affect the quality of summaries. The sixty undergraduates were randomly distributed among three treatment conditions. Each student read and summarized eight passages in one of four balanced sequences. Students in one of the three treatment conditions were specially tutored to write summaries. Students in one control treatment were asked to write summaries according to "topics" and "main points." Students in the other control treatment received no special instructions.

RESULTS

The reliability of the scores after a six-month period was found to be $r = .89$. The other test of reliability, the correlation between less stringent and more stringent scores, yielded $r = .86$.

The content validity of the index is apparent, to the degree that better summaries are those which mention more of the important ideas and use fewer words. The content validity is lacking, however, to the degree that better summaries are those with better grammar, style, or other characteristics which contribute more to aesthetic quality than to inform-

ational content, because these variables are not measured directly by the scoring procedure.

Criterion-related validity was tested by two-way analyses of variance with the scores as dependent variables. One independent variable was the judges' rating of each summary as better or not as good, and the other was the passage summarized. The main effect of the judges' rating was very significant, $p < .003$, whether the more lenient or more stringent scores were used as dependent measures. There was no significant main effect for the passages and no significant interaction.

Insert Table 1 about here

For the test of construct validity, an expected warm-up effect was found when students summarized several passages in sequence. More importantly, scores were significantly higher, as expected, for the experimentally tutored group compared to the two control groups, $p < .02$.

DISCUSSION

The results, to date, have been very promising. The scoring procedure appears to have very good reliability and validity, but its value needs to be tested by others. If others also find that the index has good reliability and validity, then teachers and researchers will have a simple procedure for scoring summaries. The index would also provide a means to evaluate the effectiveness of different instructional programs which are designed to help students grasp the main ideas of what they read or hear.

Table 1

Analysis of Variance Tables

Less Stringent Scores

Source	df	Sum of Squares	F	p <
Passage	3	.0095	1.03	.391
Rating	1	.0374	12.20	.002*
Passage*Rating	3	.0173	1.88	.151
Error	36	.1104		

More Stringent Scores

Source	df	Sum of Squares	F	p <
Passage	3	.0215	2.22	.103
Rating	1	.0344	10.66	.003*
Passage*Rating	3	.0152	1.57	.214
Error	36	.1161		

Note. With the less stringent scoring procedure, the mean for the summaries rated "better" was .171 and the mean for the summaries rated "not as good" was .113. With the more stringent procedure, the means were .149 and .089 for the "better" and "not as good" summaries, respectively.